# Predicting Team Performance Through Human Behavioral Sensing and Quantitative Workflow Instrumentation

Matthew Daggett[1], Kyle O'Brien[1], Michael Hurley[1], Daniel Hannon[1]

[1] Massachusetts Institute of Technology Lincoln Laboratory
244 Wood Street, Lexington, Massachusetts 02420, USA
{daggett, kyle.obrien, hurley, daniel.hannon}@ll.mit.edu

**Abstract.** For decades, the social sciences have provided the foundation for the study of humans interacting with systems; however, sparse, qualitative, and often subjective observations can be insufficient in capturing the complex dynamics of modern sociotechnical enterprises. Technical advances in quantitative system-level and physiological instrumentation have made possible greater objective study of human-system interactions, and joint qualitative-quantitative methodologies are being developed to improve human performance characterization. In this paper we detail how these methodologies were applied to assess teams' abilities to effectively discover information, collaborate, and make risk-informed decisions during serious games. Statistical models of intra-game performance were developed to determine whether behaviors in specific facets of the gameplay workflow were predictive of analytical performance and games outcomes. A study of over seventy instrumented teams revealed that teams who were more effective at face-to-face communication and system interaction performed better at information discovery tasks and had more accurate game decisions. [2]

**Keywords:** Humatics · Serious Games · Human-System Interaction · Instrumentation · Teamwork · Communication Analysis · Information Theory · Operations Research · Decision-Making.

## 1 Introduction

From network operations control centers to expeditionary military detachments, teams of humans interoperate with complicated systems to create complex sociotechnical enterprises. Within these enterprises, the most critical component of overall performance is that of the human, yet their contribution is often the least understood. For decades, social science has provided the foundation for the study of humans in these contexts through the observations of ethnographers and anthropologists, yet

[2] Distribution A; Public Release. This work was sponsored by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

these traditional methodologies have significant limitations. Human observation is often subjective and anecdotal and can suffer from biases and differences in interpretation. Additionally, existing tools to measure human behavior can be qualitative and are insufficient in capturing intricate intra- and inter-individual dynamics. Lastly, the collection of these data is time and human intensive and does not scale to large organizational studies. These limitations hinder the ability to draw objective conclusions and understand the parameters influencing team success. Recent technical advances in sensing and instrumentation can be used to augment human observation and enable quantitative, persistent, and objective measurements of human behavior. By jointly processing these multi-modal data, a more complete characterization of human-system interaction can be made, increasing the ability to modify behavior and improve performance. The fidelity and granularity of these data can be very revealing, and in some instances be used to predict performance in related aspects of the activities being measured.

## 2 Humatics Assessment Methodology

Over several years, we have developed a data-driven research methodology and technical framework, *Humatics*, to address the challenges outlined in Section 1 by quantitatively measuring human behavior; rigorously assessing human analytical and cognitive performance; and providing data-driven ways to improve the effectiveness of individuals and teams. Humatics incorporates three major areas of research including system-level, physiological, and cognitive instrumentation; assessment methodology and metrics development; and performance feedback and behavioral recommendation. In this paper, we describe an instantiation of this approach, shown in Fig. 1, and its application to the study of teams' abilities to effectively discover data, make sense of that data, and make decisions during a serious game.

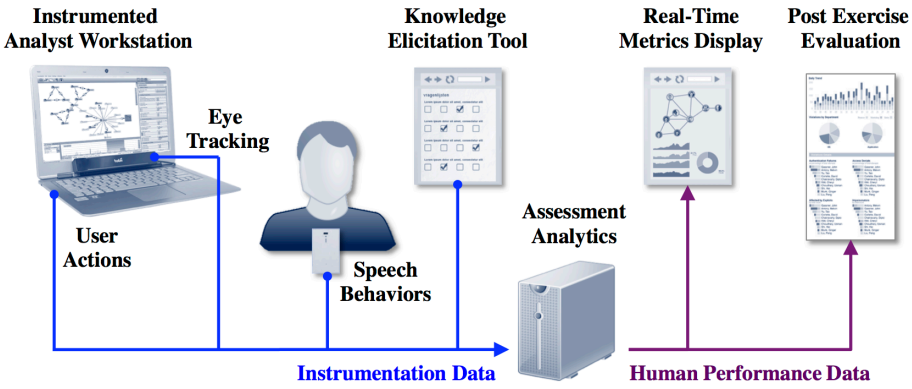


**Fig. 1.** Humatics performance assessment framework

The development of an instrumentation and data collection strategy for a given human-system research effort requires careful consideration of the specific learning

objective for the process being studied and identification of observables to be measured in order to enable insight. A measurement strategy can then be developed based on which method and phenomenology is best suited to directly or indirectly measure those observables. For this research effort, specific instrumentation modalities were chosen to augment qualitative human observations with near continuous collection to enable analysis of dynamic low-level behavioral signals.

The first element of the framework in Fig. 1 is the instrumented analyst workstation, where both system-level and physiological instrumentation are used to characterize human-system interactions. System-level instrumentation is accomplished through the insertion or enabling of software code that logs graphical user interface interaction events, queries to and transactions with databases, what data is visible to the user, and more. To add context to the data, screen recordings are continuously captured and a research-grade eye tracker is employed to detect the user's location of gaze on the screen. This physiological information is used for cross-referencing the system-level data.

The next element in Fig. 1 is cognitive instrumentation, which is used to measure behaviors associated with the cognitive processing of information. To quantify the comprehension and situational understanding of teams during scenario-based training or serious games, knowledge elicitation techniques are employed [9][15]. In addition to gaze following, the eye tracker is also used to perform pupillometry in order to noninvasively estimate human cognitive load [10].

The last framework instrumentation modality involves the use of wearable sensors, called Sociometric Badges [16], to record non-linguistic metadata of speech behaviors, body movement, and other data. Originally developed by the MIT Media Laboratory, the badges have often been employed to perform longitudinal studies of the communication patterns of large organizations. For this application, badges with modified firmware and custom post-processing software are used to increase granularity for small group dynamics within hierarchical teams.

Collected instrumentation data is processed with specialized metrics and are used for real-time diagnostic displays or post-experiment assessment. Real-time displays allow for immediate team evaluation to enable behavioral redirection, while offline post-processing supports in-depth analysis and process improvement. The team assessments in this document are an example of the latter.

## 3 Network Discovery Serious Game

In 2009, researchers at MIT Lincoln Laboratory developed a serious game to better understand how analysts use multisource textual and geospatial data to make risk-informed decisions [1][3][4]. In the game, competitive teams of varying size from 3 to 8 players analyze the scenario data to make expected decision outcomes. The teams self organize their roles and responsibilities; teams were provided with one less game client than the total number of players, generally causing hierarchies to form with one leader and the rest workers. The game scenario is based around a scripted storyboard where an organized crime network is operating in a city to incite violence (kidnappings, attacks) and then quickly disperses into the background populace of the

city. From this storyboard a probabilistic vehicle traffic model produces vehicle movements, or tracks, for the scenario vehicles the teams are tasked to find. Those tracks are embedded into realistic background tracks from the same model that simulate the normative movements of the city population. Using this combined track dataset as input, video modeling and simulation tools are used to produce a simulated airborne video dataset rendered over the city's geospatial extent for each time-step in the scenario storyboard.

Teams are given news and police reports of varying relevance to cue them to observable events in the video. Teams then analyze the video to follow suspect vehicles from overt events to their sources and destinations in order to unravel the network of facilities used by the crime organization. Teams are given 90 total minutes to accrue evidence (discovery phase) and then to codify what they know (decision phase) by identifying which facilities (sites) should be interdicted by law enforcement to disrupt the network. All scenario data are displayed, manipulated, and acted upon in a software game client that was purpose-built for this research and is instrumented for post-game analysis as described in Section 2.

## 4 Team Assessment Case Study

To assess human performance during the game, each major step of the workflow is decomposed and mapped to instrumentation data and performance metrics that characterize their behaviors. As seen in Fig. 2, three major facets of performance emerge including client interaction, information triage, and discovery and decision performance. Additionally, the performance of this entire workflow is underpinned by a team's ability to effectively organize and collaborate through face-to-face communication. In this section, a case study of four 5-member teams illustrates how system-level and physiological instrumentation can be used to better characterize a team's performance during gameplay.

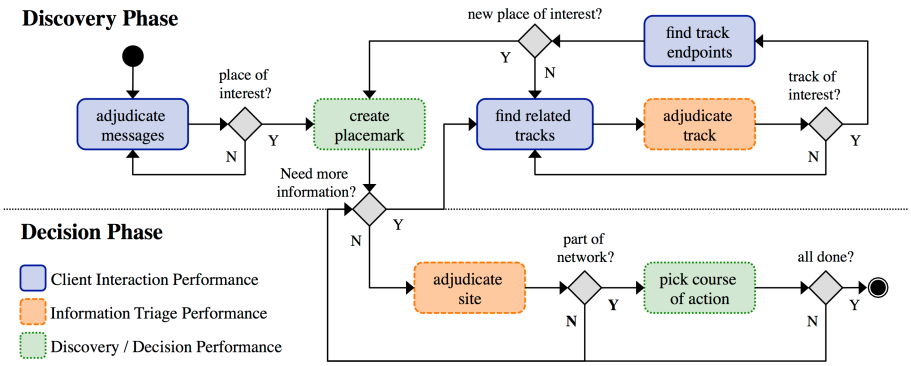


**Fig. 2.** Gameplay workflow and performance metric mapping.

### 4.1 Game Client Interaction Performance

Software instrumentation built into the game client records various user interactions both on-demand and at specific intervals. These data can be used to understand macro behaviors like the volume or rate of interactions with specific tools in the client. For example, in the game teams use placemarks, geo-spatial annotations, to codify the site discoveries they have made and to later code their courses of action (decisions). By recording placemark creation and modification attributes we can quantify team analytical behaviors in the workflow as a function of time. These data can also be used to analyze micro behaviors, such as the geospatial data currently being viewed by the user, known as the viewport [1]. Viewport data are recorded each second and include the current time of gameplay, the time in the scenario being displayed, and the geospatial bounding-box of the video footprint in the map tool. An example of viewport instrumentation is shown in **Fig. 3**.

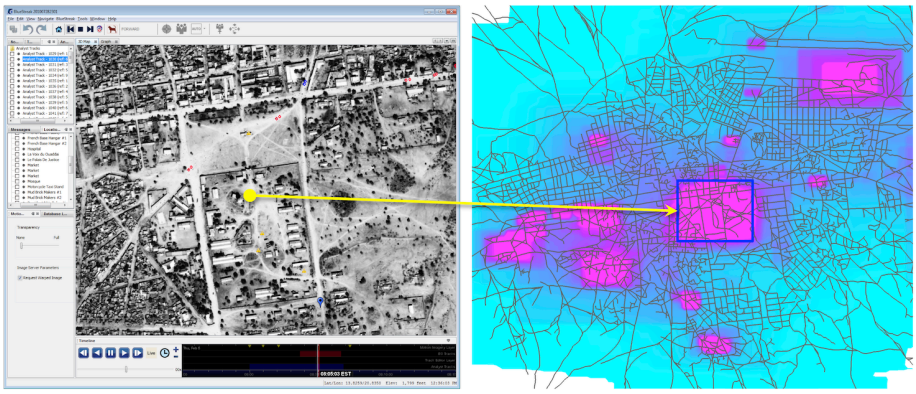

**Fig. 3.** An illustration of viewport instrumentation. The game client (left) is viewing a portion of the scenario video, whose viewport extents are represented by the blue box on the heat-map (right), as indicated by the yellow arrow. Viewport heat-maps can be used to understand a team's geospatial analysis strategy.

### 4.2 Scenario Information Triage Performance

After the viewport data are logged as described in Section 4.1, they are correlated with the scenario ground truth and processed using specialized information theoretic metrics [1][2] to determine which relevant (scenario crime network) and irrelevant (background population) tracks or sites are being viewed at each scenario time-step. A graphical representation of the scenario and background track information is shown in Fig. 4, which illustrates the teams' ability to effectively triage vehicle track data. If players are properly interpreting the information in the report messages they should focus only on the red scenario vehicle tracks and not the yellow background population tracks. As shown in Fig. 4, Teams 3 and 4's performance plateaus as the scenario evolves, whereas Teams 1 and 2 continue to find and analyze more relevant (red) scenario tracks throughout gameplay.

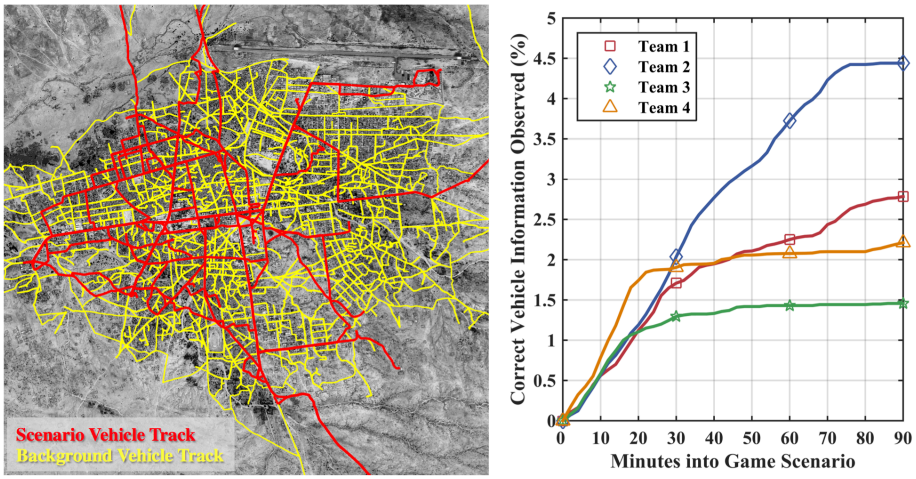


**Fig. 4.** Team vehicle triage performance. The left plot shows the extents of all vehicle track data in the game, with the red lines denoting tracks associated with the criminal network vehicles and the yellow tracks of background population tracks. The right plot shows the team triage performance, with the y-axis representing the percentage of total red tracks observed in the video and the x-axis representing the number of minutes elapsed since the start of the game.

Similarly, Fig. 5 illustrates the teams' ability to effectively triage video of site-related activities. As shown in the figure, Teams 1 and 2 spent substantially more effort observing scenario site information as compared to Teams 3 and 4. In many cases teams, spend a lot of time analyzing sites but ultimately chose an incorrect action or take no action at all.

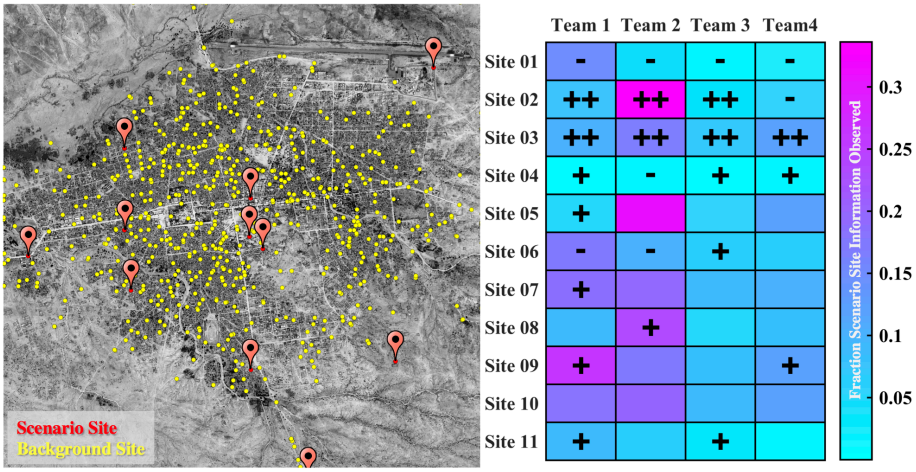


**Fig. 5.** Team site triage performance. The left plot shows the scenario sites to be discovered, as annotated with the red icons, and the background sites, denoted with a yellow dot. The right plot shows the performance of the teams at accumulating information at each of the scenario sites, as indicated by the fill color of each box and color bar scale. Decision outcomes for sites are also plotted (right), with a + or - representing a correct or incorrect decision respectively.

### 4.3 Team Discovery and Decision-Making Performance

Because the scenario was constructed to have the crime network activities completely separated from the background activity, the game can be analyzed from the perspective of signal detection theory. Essentially, the teams are considered to be detectors of criminal network activity in that they are attempting to extract these signals from the noise of the normal activities of the rest of the population [3]. The Receiver Operating Characteristics (ROC) measurements of detection theory can be used to assess the teams' performance, as shown in Fig. 6.

Results from two different tasks are plotted: the discovery of scenario sites as measured by team placemarks at those sites, and the declaration of scenario sites which are the subset of the total placemarks that are given a course of action decision. Decision actions are directly related to the teams' comprehension of the scenario storyboard and their confidence in that understanding. For example, it can be seen that Team 2 had placemarks on 100% of the crime network sites, but only had the confidence to declare 30% of those sites. They also declared sites not part of the network, resulting in a 0.2% probability of false declaration. Team 4 had similar discovery performance as Team 1 and 0% probability of false declaration. Team 1 had the highest detection probability, but at the expense of more false declarations.

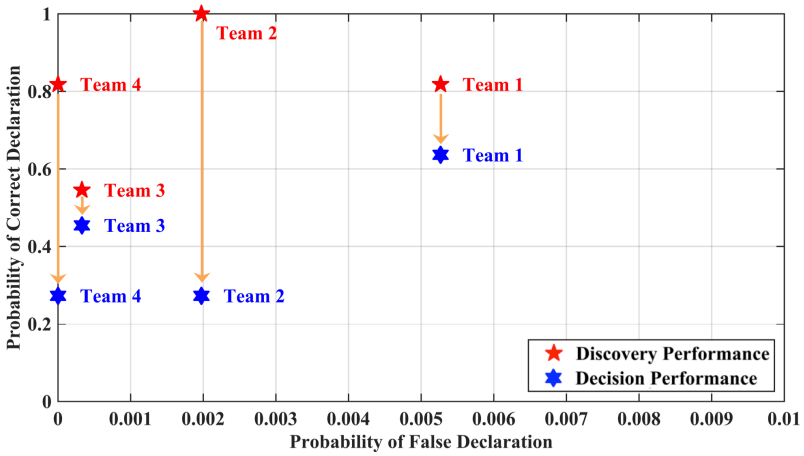


**Fig. 6.** Team discovery and decision ROC plot. The y-axis represents the *Probability of Correct Declaration*, or the fraction of correct sites found and acted upon by the teams. The x-axis represents the *Probability of False Declaration*, or the ratio of incorrect sites declared divided by the total possible discoverable sites. The blue icons represent decision performance for sites that were declared to be associated with the network. The red icons show the fraction of all sites that were correctly discovered before the course of action selection process.

### 4.4 Team Verbal Communication Performance

Face-to-face communication is known to be a key factor in overall team performance for highly cooperative tasks [11][12][13][14]. Traditional methods to characterize these communications have largely focused on speech content, however

more recent methods center on the collection of non-linguistic speech features that enable the characterization of team dynamics without having to analyze the linguistic content of a team's utterances [12].

To collect speech metadata, Sociometric Badges are given to each player during gameplay. The badges continuously record the time, duration, and identity of each player's speech, and post-processing software provides measurements of when players spoke alone, when speech overlapped with another player, which players were listening, and when players were silent. These data naturally form a directed graph of communication between players as shown in Fig. 7. For simplicity, only Teams 1 and 2 are shown.

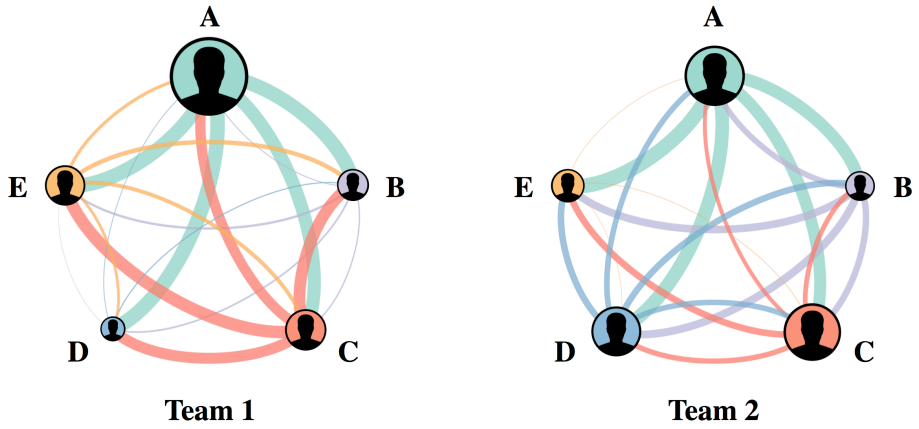


**Fig. 7.** Face-to-Face communication network graphs. Vertices represent players and edges represent directed communication from one player to another. Vertex size is proportional to total participation for a player, edge thickness is proportional to directed speech time to each teammate, and edge color indicates directionality by matching the source vertex color.

Previous studies of face-to-face communication behaviors of small teams in a collaborative setting have found that balanced participation and speaking time along with increased turn-taking are associated with better team performance [8]. In Fig. 7, Team 1 players A and C are dominating the conversation as seen by their edge thickness, while the rest of the players are less engaged with lower participation (smaller vertices) and less speaking time (thinner edges). Conversely, Team 2 has a much more balanced distribution of both speaking time and participation than Team 1, with player A acting in the role of team leader. Analysis of group influencers and team role estimation using these data is a promising area of active research [7].

For deeper insight into the communication network, a Social Network Analysis approach to characterizing player interaction is explored. By computing the directed, normalized Closeness Centrality of each player [5], an estimate of the connectedness of players can be derived. Larger centrality magnitudes indicate a player's graph "closeness" to all other players. One useful application of this measure is to inspect the time-varying behavior of player centrality [6] during game play, as shown in Fig. 8. In the figure, the visual representation of Teams' 1 and 2 closeness centrality can be very useful for identifying team dynamic attributes, such as leader emergence. For Team 1 we see the same effect of players A and C communication dominance as seen

in Fig. 7. For Team 2 we see the clear emergence of player A as the leader of the team during the discovery phase of the game with A's centrality reducing towards the end as the team moved into the collective decision-making phase of the game.

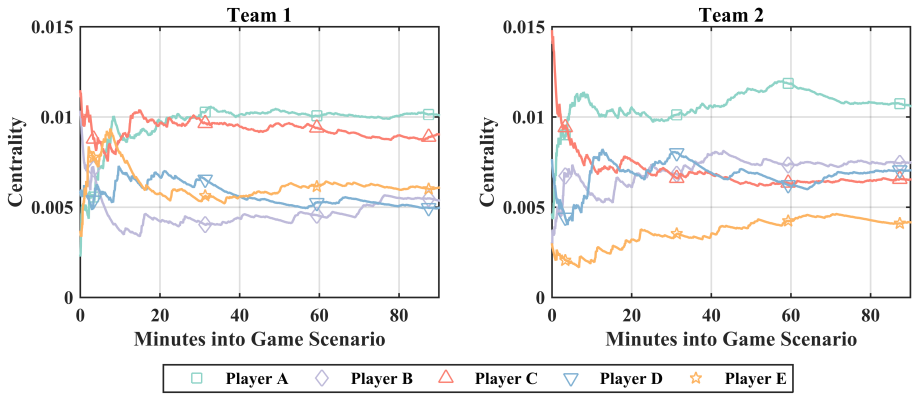


**Fig. 8.** Time-varying player communication centrality. In both teams, Player A is considered the leader and transitions to gain the highest *Centrality* midway through the game. Qualitative observations during gameplay support these findings.

In addition to Social Network Analysis, Recurrent Pattern Analysis was also performed using the Sociometric Badge data. First, speech patterns are coded into symbols according to various speech behaviors and are then analyzed as a time series [17]. The strength of the recurrent structure within these code sequences is called determinism (DET). In a strict turn taking situation the DET will be high (near 100%) as the conversation is highly structured. In a situation with random speech intervals, the DET will be low (close to 0%) indicating that the conversation is highly unstructured. DET scores were comparable for the four teams, with local maxima near 60% and local minima near 30%. There were fluctuations in the values over time, indicating that the structure of the communication ebbed and flowed throughout gameplay. Further analysis showed a high correlation between DET magnitude and the percentage of time an individual spoke while all others listened ($r=0.47$-$0.53$, $p<0.001$), suggesting that in part, structure occurs, even in a complex team setting with five participants, when individuals speak and others listen.

### 4.5 Total Team Performance

Sections 4.1 through 4.4 demonstrated how performance is quantitatively measured at several points in the overall game workflow, however combining these metrics into a single total performance measure warrants careful consideration. Qualitatively, Teams 1 and 2 excelled at communication, triage, and site discovery, but had more false declarations than Teams 3 and 4. Conversely, Teams 3 and 4 did not observe as much information or discover as many sites, but were very accurate in adjudicating what they found. Team 2 ultimately was the winner of the four-team competition and had the best overall performance and game score.

## 5 Predicting Team Performance

When assessing teams' analytical and decision-making performance, common questions arise regarding how performance in one facet of a decision process affects the performance of either subsequent processes or the aggregate overall process. Section 4 illustrates that the collected measurements enable detailed insight about individual facets of performance; however, we wanted to take this a step further to determine whether behaviors in specific facets of the intra-game workflow were predictive of analytical performance or games outcomes. To approach this, we processed data collected over several years of employment of the game, encompassing 71 different teams comprised of over 350 unique players. For all of the 71 teams, system instrumentation data were recorded. For a subset of 15 of those teams, face-to-face communication data were also collected.

Robust linear regression analyses are used to statistically estimate how predictive various facets of intra-game performance are with respect to workflow processes. For each model, residual analysis, significance testing, and other regression diagnostics are performed, and their results are included in parentheticals with each prediction finding. In undertaking this analysis, we want to address three overarching research propositions, as follows.

### 5.1 Client Interaction Effectiveness

The first proposition investigated was whether more effective interaction with the game software client led to better game performance. From our analysis, we found that teams (N=71) who had higher usage across all analytic functions of the game client discovered more total sites ($p<0.001$, $R^2=0.25$) and had a higher probability of correct site discovery ($p<0.001$, $R^2=0.20$). The effect was even more pronounced for the functions of the game client associated with the frequency that players submitted space-time queries for track data and its correlation with increased site discovery ($p<0.001$, $R^2=0.43$). Higher total game client interaction was also associated with more effective observations of scenario site information ($p<0.001$, $R^2=0.21$) and track information ($p=0.006$, $R^2=0.20$). Essentially teams who were more effective at interacting with the functions of the game client observed more relevant scenario information and found more correct sites.

### 5.2 Information Triage Effectiveness

The second proposition investigated was whether discovery of more scenario information led to better game outcomes. From our analysis, we found that teams (N=71) who observed more relevant scenario site information ($p=0.002$, $R^2=0.15$) and track information ($p=0.006$, $R^2=0.13$) also scored higher in game outcome. The overall game score is a metric that takes into account several aspects of how well the players perform but it also encapsulates the confidence of their decisions (courses of action strength) and reflects the overall strategy for how they decide to approach the game (aggressive to risk-averse).

### 5.3 Team Communication Effectiveness

The third proposition investigated was whether teams who communicate more effectively have higher game performance. Our analysis found that teams (N=15) who communicated more (total time) throughout the exercise also observed more relevant scenario site information ($p=0.006$, $R^2=0.51$) and track information ($p=0.001$, $R^2=0.61$). Additionally, teams who had higher participation (frequency of communication) from all of their members throughout the game also observed more relevant scenario site information ($p=0.002$, $R^2=0.60$) and track information ($p=0.010$, $R^2=0.46$). Lastly, teams who communicated more (total time) throughout the exercise also made better decisions on the most challenging sites to adjudicate ($p=0.007$, $R^2=0.47$). This agrees with qualitative observations of teams during the decision-making process regarding the total engagement and participation of the full team. Team centrality metrics, as shown Section 4.4, did not have a significant association with other aspects of team performance, and warrants further investigation.

## 6 Conclusion

In this paper we have shown that Humatics can be used to improve the quantitative study and objective assessment of human analytic and decision-oriented processes, and we detailed its application to measuring a team's ability to effectively discover information, collaborate, and make decisions during a serious game. Additionally, we have demonstrated that data collected about intra-game workflow process can be used to predict subsequent game outcomes and other performance attributes. While we described one specific instantiation of Humatics, the framework has broad applicability towards the optimization of a wide range of complex sociotechnical enterprises. Future research will study the combination of multiple sources of heterogeneous instrumentation data and identifying their relationships to other aspects of human behavior and performance.

## References


1. Daggett, M., O'Brien, K., Hurley, M.: An Information Theoretic Approach for Measuring Data Discovery and Utilization During Analytical and Decision-Making Processes, Games and Learning Alliance: Fourth International Conference, GALA, Rome, Italy (2015)
2. Kao, E.K., Daggett, M.P., Hurley, M.B.: An Information Theoretic Approach for Tracker Performance Evaluation, IEEE 12th International Conference on Computer Vision (ICCV), Kyoto, Japan (2009)
3. Won, J.C.: Influence of Resource Allocation on Teamwork and Performance in an Intelligence, Surveillance, and Reconnaissance (ISR) "Red/Blue" Exercise Within Self-Organizing Teams, PhD Thesis, Tufts University (2012)
4. Won, J.C., Condon, G.R., Landon, B.R., Wang, A.R., Hannon, D.J.: Assessing team workload and situational awareness in an Intelligence, Surveillance, and Reconnaissance (ISR) simulation exercise, IEEE First International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA) (2011)

5. Linton C. Freeman: Centrality in networks: I. Conceptual clarification. Social Networks pp. 1:215-239 (1979)
6. Ara, K., et al. Sensible Organizations: Changing Our Business and Work Styles through Sensor Data. Journal of Information Processing, Vol. 16, 1-12, April (2008)
7. Dong, W., et al. Using the Influence Model to Recognize Functional Roles. MIT Media Laboratory Technical Note 609. Ninth Int'l Conf on Multimodal Interfaces, Nov 12-15, Nagoya Japan (2007)
8. Dong, W., Lepri, B., Kim, T., Pianesi, F., and Pentland, A. Modeling Conversational Dynamics and Performance in a Social Dilemma Task. 5th International Symposium on Communications, Control, and Signal Processing. Povo-Trento, Italy, May (2012)
9. Endsley, M., Situation awareness misconceptions and misunderstandings. Journal of Cognitive Engineering and Decision Making, 9(1), pp. 4 – 32 (2015)
10. Klinger, J., Kumar, R., & Hanrahan, P., Measuring the task-evoked pupillary response with a remote eye tracker. Proceedings of the Eye Tracking Research and Applications Symposium, ETRA,2008, Savannah, GA, USA, pp. 69 – 72 (2008)
11. Apedoe, X.S., Mattis, K.V., Rowden-Quince, B., & Schunn, C.D. Examining the role of verbal interaction in team success on a design challenge. Learning in the Disciplines: ICLS 2010 Conference Proceedings – 9th International Conference of the Learning Sciences, Chicago, IL, pp. 596 – 603 (2010)
12. Strang, A, Horwood, S., Best, C., Funke, G, Knott, B.A., and Russell, S. M. Examining temporal regularity in categoritcal team communication using Sample Entropy. Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting, HFES 2012, pp. 473 - 477 (2012)
13. Whitaker, L.A. & Peters,L.J.. Communication between crews: The effects of speech intelligibility on team performance. Proceedings of the Human Factors and Ergonomics Society, Seattle, WA, pp. 630-634 (1993)
14. Andres, H.P. The impact of communication medium on virtual team group process. Information Resources Management Journal, 19(2), pp. 1-17 (2006)
15. Salmon, P., Stanton, N., Walker, G., & Green, D. Situation awareness measurement: A review of applicability for C4i environments. *Applied ergonomics*, *37*(2), 225-238 (2006)
16. Olguín-Olguín, D. and Pentland, A., Sensor-based organisational design and engineering. *International Journal of Organisational Design and Engineering, 1*(1-2), pp. 69-97 (2010)
17. Gorman, J.C., Cooke, N.J., Amazeen, P.G., & Fouse, S.. Measuring patterns in team interaction sequences using a discrete recurrence approach. Human Factors, 54(1), pp. 503 – 517 (2012)